Acta Univ. Palacki. Olomuc., Fac. rer. nat.,
Mathematica **50**, 2 (2011) 111–119

Concept of Data Depth and Its Applications*

Ondřej VENCÁLEK

Department of Mathematical Analysis and Applications of Mathematics
Faculty of Science, Palacký University
17. listopadu 12, 771 46 Olomouc, Czech Republic
e-mail: ondrej.vencalek@upol.cz

Dedicated to Lubomír Kubáček on the occasion of his 80th birthday

(Received March 31, 2011)

Abstract

Data depth is an important concept of nonparametric approach to multivariate data analysis. The main aim of the paper is to review possible applications of the data depth, including outlier detection, robust and affine-equivariant estimates of location, rank tests for multivariate scale difference, control charts for multivariate processes, and depth-based classifiers solving discrimination problem.

Key words: data depth, nonparametric multivariate analysis, applications, rank

2010 Mathematics Subject Classification: 62G05, 62G15, 60D05, 62H05

1 Introduction

Data depth is an important concept of nonparametric approach to multivariate data analysis. It provides one possible way of ordering the multivariate data. We call this ordering a central-outward ordering. Basically, any function which provides a "reasonable" central-outward ordering of points in multidimensional space can be considered as a depth function. This vague understanding of the notion of depth function led to the variety of depth functions, which have been introduced ad hoc since 1970s. The formal definition of a depth function was formulated by Zuo and Serfling in 2000 [8].

*Supported by the grant GAUK B-MAT 150110.

The most widely used depth function is the halfspace depth function. The halfspace depth of a point $\vec{x}$ in $\mathbb{R}^d$ with respect to a probability measure P is defined as the minimum probability mass carried by any closed halfspace containing $\vec{x}$, that is

$$D(\vec{x}; P) = \inf_H \left\{ P(H) \colon H \text{ a closed halfspace in } \mathbb{R}^d \colon \vec{x} \in H \right\}.$$

The halfspace depth is sometimes called location depth or Tukey depth, as it was first defined by Tukey in 1975 [7]. The halfspace depth is well defined for all $\vec{x} \in \mathbb{R}^d$. Its sample version (empirical halfspace depth), defined on a random sample $\vec{X}_1, \ldots, \vec{X}_n$ of the distribution P, is defined as the halfspace depth for the empirical probability measure P_n. This definition is very intuitive and easily interpretable. Moreover, there are many desirable properties of the halfspace depth, which made this depth function very popular and widely used. In particular, the halfspace depth is affine invariant and has all the other desirable properties stated in the general definition of depth function by Zuo and Serfling.

The notion of *rank* is crucial in many applications. Consider a d-dimensional probability distribution P and a random sample $\vec{X}_1, \ldots, \vec{X}_n$ from this distribution. (The empirical probability measure based on the sample is again denoted by P_n). For any point $\vec{x} \in \mathbb{R}^d$ we define

$$r_P(\vec{x}) = P(D(\vec{X}; P) \leq D(\vec{x}; P) \mid \vec{X} \sim P) \tag{1}$$

and

$$r_{P_n}(\vec{x}) = \# \left\{ \vec{X}_i \colon D(\vec{X}_i; P_n) \leq D(\vec{x}; P_n),\ i = 1, \ldots, n \right\} / n. \tag{2}$$

2 Outlier detection—a bagplot

Rousseeuw et al. [6] proposed a bivariate generalization of the univariate boxplot, so called bagplot. They used the halfspace depth to order the data, but other depth functions might be used as well. The bagplot consists of

- the deepest point (the point with maximal depth),
- the bag, that is the central area, which contains 50 % of all points; the bag is usually dark colored,
- the fence, which is found by magnifying the bag by a factor 3; the fence is usually not plotted; observations outside the fence are flagged as outliers,
- the loop, which is an area between the bag and the fence; usually light coloured.

The bagplot procedure is available in `R` library `aplpack`. As an example, we plot the bagplot of the car data of Chambers and Hastie that are available in library `rpart`. Figure 1 displays car weight and engine displacement of 60 cars. Five outliers were detected.

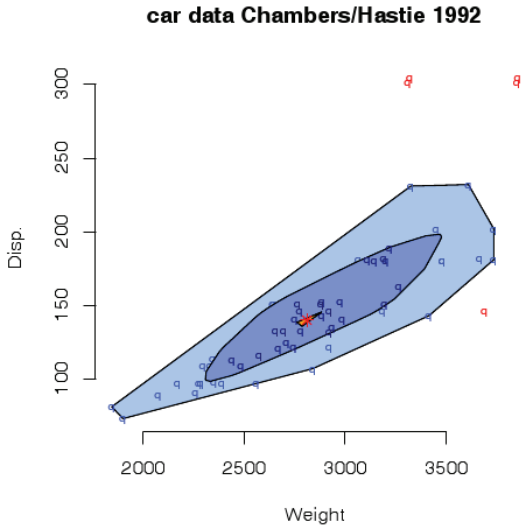

Figure 1: An example of bagplot.

3 Affine-equivariant and robust estimates of location

Donoho and Gasko [1] have shown that two basic location estimators based on the halfspace depth, the deepest point and the trimmed mean (with trimming based on the halfspace depth), are both affine equivariant and robust (in the sense of the high breakdown point). The combination of these two properties is quite rare in multivariate statistics. The most important results are summarized in the next theorem:

Theorem 1 *Let $\vec{X}_1, \ldots, \vec{X}_n$ be a sample determining empirical version P_n of an absolutely continuous distribution P on $\mathbb{R}^d$, with $d > 2$. Assume data be in a general position (no ties, no more than two points on any line, three in any plane, and so forth).*

Consider the deepest point $T_*(P_n) = \arg\max_{\vec{x}} D(\vec{x}, P_n)$ and α-trimmed mean $T_\alpha(P_n) = Ave(\vec{X}_i \colon D(\vec{X}_i; P_n) \geq n\alpha)$, the average of all points whose depth is at least $n\alpha$.

Denote $\beta := \arg\max_{\vec{x}} D(\vec{x}; P)$ ($\beta = 1/2$ if P is centrally symmetric). Then

1. *The breakdown point of $T_*(P_n)$ is greater or equal to $1/(d+1)$. It converges almost surely to $1/3$ as $n \to \infty$ if P is centrally symmetric.*

2. *For each $\alpha \leq \beta/(1+\beta)$, $T_\alpha(P_n)$ is well defined for sufficiently large n and its breakdown point converges almost surely to α.*

4 Rank tests for multivariate scale difference

Liu and Singh [4] combined ranks based on data depth with well-known one-dimensional nonparametric procedures to test scale difference between two or more distributions.

Consider two d-dimensional distributions P_1 and P_2, which possibly differ in dispersion only. Denote $\vec{X}_1, \ldots, \vec{X}_{n_1}$ a random sample from P_1 and $\vec{Y}_1, \ldots, \vec{Y}_{n_2}$ a random sample from P_2. Denote the combined sample as $\left\{\vec{W}_1, \ldots, \vec{W}_{n_1+n_2}\right\} \equiv \left\{\vec{X}_1, \ldots, \vec{X}_{n_1}, \vec{Y}_1, \ldots, \vec{Y}_{n_2}\right\}$ and denote $P_{n_1+n_2}$ the empirical distribution function based on the combined sample.

We want to test the hypothesis H_0 of equal scales against the alternative that P_2 has larger scale in the sense that the scale of P_2 is an expansion of the scale of P_1. If the scale of P_2 is greater, then obviously observations from the second distribution tend to be more outlying than the observations from P_1. Consider the sum of the non-normalized ranks for the sample from P_2:

$$R\left(Y_1, \ldots, Y_{n_2}\right) = (n_1 + n_2) \sum_{i=1}^{n_2} r_{P_{n_1+n_2}}(Y_i).$$

Now we proceed as in the case of testing for a (negative) location shift in the univariate setting. This leads us to the Wilcoxon rank-sum procedure. When n_1 and n_2 are sufficiently large, we can rely on asymptotic behaviour of the test statistic (assuming null hypothesis):

$$R^* = \frac{R(\vec{Y}_1, \ldots, \vec{Y}_{n_2}) - [n_2(n_1 + n_2 + 1)/2]}{[n_1 n_2 (n_1 + n_2 + 1)/12]^{1/2}} \xrightarrow{D} N(0, 1),$$

and hence we reject H_0 if $R^* \leq \Phi^{-1}(\alpha)$, where $\Phi^{-1}(\alpha)$ is the α-quantile of the standard normal distribution.

We can proceed similarly when considering more than two (say $K > 2$) distributions. We test the hypothesis that the underlying distributions are identical against the alternative that the scales of these distributions are not all the same, in the sense of scale contraction. Construction of the test follows the idea of the well-known Kruskal-Wallis test. Let $\bar{R}_i$ denote the average of non-normalized ranks (based on data depth) of the observations from the i-th sample in the combined sample. The total number of all observations in combined sample (from all K samples) is N. Under the null hypothesis, it holds:

$$T = \frac{12}{N(N+1)} \sum_{i=1}^{K} \left(n_i \bar{R}_i^2\right) - 3(N+1) \xrightarrow{D} \chi^2_{K-1}.$$

We reject the null hypothesis at an approximate level α if $T \geq \chi^2_{K-1}(1-\alpha)$, where $\chi^2_{K-1}(1-\alpha)$ is the $(1-\alpha)$ quantile of a chi-squared distribution with $(K-1)$ degrees of freedom.

There is a simple graphical tool developed by Liu, Parelius and Singh (see [5]) to visualize difference in scales of multivariate distributions. They defined

a *scale curve* as a plot of $p \in (0,1)$ versus volume of C_p - the p-th central region (the *p-th central region* C_p is defined as the smallest region enclosed by depth contours to amass probability p, that is $C_p = \bigcap_t \{R(t)\colon \mathrm{P}(R(t)) \geq p\}$, where $R(t) = \left\{\vec{x} \in \mathbb{R}^d\colon D(\vec{x}; P) > t\right\}$). The sample scale curve, based on random sample $\vec{X}_1, \ldots, \vec{X}_n$, plots volumes of the convex hulls containing $\lceil np \rceil$ most central points versus p. By plotting scale curves for compared distributions in one plot, the difference in scales can be easily visualized.

The following example should illustrate the methodology. We simulated 250 points from bivariate $N(\vec{0}, \boldsymbol{I})$ distribution and the same number of points from $N(\vec{0}, 2\boldsymbol{I})$ ($\boldsymbol{I}$ denoting 2×2 identity matrix). The test statistic was $R^* = -2,14$, which is less than $\Phi^{-1}(0,05) = -1,64$. We thus (correctly) reject the null hypothesis of identical distributions. The difference in dispersions can be seen in Figure 2.

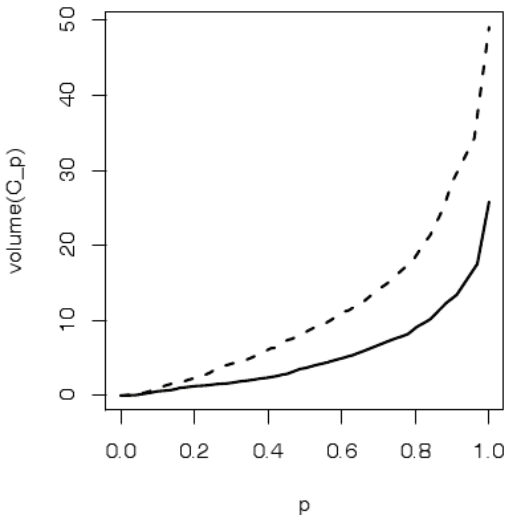

Figure 2: Empirical scale curves based on samples of 250 points from $N(\vec{0}, \boldsymbol{I})$ (solid line) and from $N(\vec{0}, 2\boldsymbol{I})$ (dashed line).

5 Control charts for multivariate processes

Liu [3] used the concept of data depth to introduce control charts for monitoring processes of multivariate quality measurements. The idea is to work with ranks of the multivariate measurements (based on data depth) rather than with multivariate measurements themselves.

Let G denote the prescribed d-dimensional distribution (if the measurements follow the distribution G, the process is considered to be in control). G is either known or it can be estimated: G_n denotes its empirical version, based on n observations. Let $\vec{X}_1, \vec{X}_2, \ldots$ be the new observations from the considered

process. They follow some distribution F. Our task is to test the null hypothesis $H_0\colon F \equiv G$ against the alternative H_A : there is a location shift or a scale increase from G to F.

The test is based on ranks $r_G(\vec{X}_1), r_G(\vec{X}_2) \ldots$ (or $r_{G_n}(\vec{X}_1), r_{G_n}(\vec{X}_2), \ldots$ if G needs to be estimated). Under the null hypothesis, it holds:

1. $r_G(\vec{X}) \sim U[0,1]$,
2. $r_{G_n}(\vec{X}) \xrightarrow{D} U[0,1]$, provided that $D(\cdot; G_n) \to D(\cdot; G)$ uniformly as $n \to \infty$.

The uniform convergence of $D(\cdot; G_n)$ holds for example for halfspace depth if G is absolutely continuous. The expected value of $r_G(\vec{X})$ is thus 0.5. Small values correspond to a change in the process. A so-called lower control limit is thus equal to α (typically 0.05). Values $r_G(\vec{X}_i) < \alpha$ signalize a possible quality deterioration.

Similarly as Liu [3], we can demonstrate the procedure on simulated data. Let the prescribed distribution G be a bivariate standard normal distribution. Firstly, we generate 500 observations from this distribution to get a sample version G_n (we consider G to be unknown to mimic some real applications). Subsequently, we generate new observations—40 observations from bivariate standard normal distribution (process in control) and next 40 observations from bivariate normal distribution with shifted mean $(2,2)^T$ and both scales doubled. The control chart is shown in Figure 3.

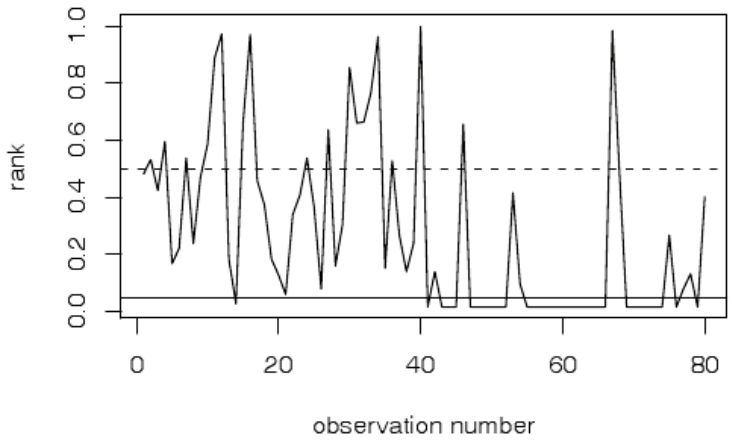

Figure 3: Control chart for multivariate process.

There is one so called false alarm in the first half of observations. The out-of-control status in the second half of observations is correctly detected 30 times (from 40 observations). The change is apparent from the chart.

Liu called this type of control chart the r chart. She also proposed multivariate versions of Shewhart chart (Q chart) and CUSUM chart (S chart).

6 Depth-based methods of discrimination

During the last ten years quite a lot of effort has been put into development of a nonparametric approach to the discrimination problem, which uses the methodology of data depth.

Recall the nub of the discrimination problem. Consider $k \geq 2$ groups of objects. Each object can be represented by $d \in \mathbb{N}$ numerical characteristics. Each group of objects is characterized by the distribution of the numerical characteristics of its members. We denote these distributions $P_1, \ldots, P_k$. The distributions are unknown. In what follows we assume the distributions to be absolutely continuous. Consider further k independent random samples $\vec{X}_{i,1}, \ldots, \vec{X}_{i,n_i}$, $i = 1, \ldots, k$, from distributions $P_1, \ldots, P_k$. These random samples (known as the training set) provide the only available information on the considered distributions. Any vector $\vec{x} \in \mathbb{R}^d$, representing an object not included in the training set, is considered to be a realization of a random vector from one of the distributions $P_1, \ldots, P_k$, but it is unknown from which of them. There is a need to estimate to which group the object belongs. The goal is to find some general rule, which allocates an arbitrary d-dimensional real vector to one of the considered distributions (groups). The rule (known as classifier) has a form of some function $d\colon \mathbb{R}^d \to \{1, \ldots, k\}$.

Probably the most widely used classifier based on data depth is a so-called maximal depth classifier. It is based on a simple idea of assigning a new observation (represented by vector $\vec{x}$) to the distribution, with respect to which it has maximal depth. An arbitrary depth function can be used, i.e.

$$d(x) = \arg \max_{j=1\ldots,k} D(\vec{x}; P_j). \tag{3}$$

Since the theoretical depth is usually unknown, empirical version based on the data from the training set is used:

$$d(x) = \arg \max_{j=1\ldots,k} D(\vec{x}; \widehat{P}_j), \tag{4}$$

where $D(\vec{x}; \widehat{P}_j)$ is a depth of $\vec{x}$ with respect to empirical distribution of the j-th distribution, which is based on the appropriate points from the training set $(\vec{X}_{j,1}, \ldots, \vec{X}_{j,n_j})$.

A detailed inspection of the method is provided in a paper by Ghosh and Chaudhuri [2]. The maximal depth classifier is known to be asymptotically optimal (it has the lowest possible average misclassification rate) in certain situations. Ghosh and Chaudhuri showed asymptotical optimality of the classifier for a very special case, assuming that the considered distributions:

- are elliptically symmetric with the density functions strictly decreasing in every direction from their centers of symmetry,
- differ only in location (have equal dispersions and are of the same type),
- have equal prior probabilities.

In addition, the used depth function must also satisfy some conditions. Ghosh and Chaudhuri formulated the following optimality theorem:

Theorem 2 *Suppose $P_1, \ldots, P_k$ are elliptically symmetric distributions with densities $f_i(\vec{x}) = g(\vec{x} - \vec{\mu}_i)$, $i = 1, \ldots, k$, where g satisfies: $g(c\vec{x}) < g(\vec{x})$ for every $\vec{x}$ and constant $c > 1$. Consider equal prior cases. When using halfspace, simplicial, majority, or projection depth, the average misclassification rate of an empirical depth-based classifier (4) converges to the optimal Bayes risk as* $\min(n_1, \ldots, n_k) \to \infty$.

The maximal-depth classifier is not optimal when the considered distributions differ in dispersion. This fact can cause serious problems even in a very simple situation: consider, for example, two bivariate normal distributions with equal prior probabilities $P_1 = \mathrm{N}\left((0,0)^T, 4\boldsymbol{I}\right)$, and $P_2 = \mathrm{N}\left((1,0)^T, \boldsymbol{I}\right)$, where $\boldsymbol{I}$ denotes 2×2 identity matrix. Denote the new observation $\vec{x} = (x_1, x_2)^T$. In this case the optimal Bayes rule has the following form: $d(\boldsymbol{x}) = 2$ iff $(x_1 - 4/3)^2 + x_2^2 < 4/9 + 16/3 \ln 2$. Expected misclassification rate for the group 1 is about 0.3409, for group 2 it is about 0.1406, hence the optimal Bayes risk is about 0.2408. The theoretical maximal depth classifier, which is equivalent to the classifier minimizing Mahalanobis distance, has the form: $d(\boldsymbol{x}) = 2$ iff $(x_1 - 4/3)^2 + x_2^2 < 4/9$. Expected misclassification rate is 0.0435 for group 1 and 0.8104 for group 2, yielding the average misclassification rate of about 0.4270, which is much higher than the optimal Bayes risk. (The expected misclassification rates were enumerated by the numeric integration of densities).

As we can see from the example above, the class of problems that can be satisfactorily solved using the classifier (4) is quite narrow. The problem of maximal-depth classifier arises from the discrepancy between the depth and the density function. The optimal Bayes classifier is based on density function. While the depth function is affine invariant, the density function does not have this property. More sophisticated classifiers are needed to overcome this problem.

7 Conclusion

The concept of data depth provides a useful tool for nonparametric multivariate statistical inference. Ordering (and ranks) based on data depth provides a basis for many nonparametric multivariate procedures like outlier detection, estimation of some basic random vector characteristics, testing for multivariate scale difference, construction of control charts for multivariate processes, or construction of classifiers for solving discrimination problem.

References

[1] Donoho, D. L., Gasko, M.: *Breakdown properties of location estimates based on halfspace depth and projected outlyingness.* Annals of Statistics **20** (1992), 1803–1827.

[2] Ghosh, A. K., Chaudhuri, P.: *On Maximum Depth and Related Classifiers.*. Scandinavian Journal of Statistics **32** (2005), 327–350.

[3] Liu, R. Y.: *Control charts for multivariate processes.*. Journal of the American Statistical Association **90** (1995), 1380–1387.

[4] Liu, R. Y., Singh, K.: *Rank tests for multivariate scale difference based on data depth.* In: Liu R. Y., Serfling R., Souvaine D. L. (eds.) DIMACS; Robust Multivariate Analysis, Computational Geometry and Applications *American Mathematical Society*, 2006, 17–34.

[5] Liu, R. Y., Parelius, J. M., Singh, K.: *Multivariate analysis by data depth: Descriptive statistics, graphics and inference (with discussion).* Annals of Statistics **27** (1999), 783–858.

[6] Rousseeuw, P. J., Ruts, I., Tukey, J.: *The bagplot: a bivariate boxplot.* The American Statistician **53** (1999), 382–387.

[7] Tukey, J.: *Mathematics and picturing data.* Proceedings of the 1975 International Congress of Mathematics **2** (1975), 523–531.

[8] Zuo, Y., Serfling, R.: *General notion of statistical depth function.* Annals of Statistics **28** (2000), 461–482.